Categorization of Free-Text Problem Lists : an Effective Method of Capturing Clinical Data

Julian Zelingher, M.D., M.Sc., David M. Rind, M.D., M.S., Enrique Caraballo, B.S.,
Mark S. Tuttle, D.D., Nels E. Olson and Charles Safran, M.D., M.S.
Beth Israel Hospital, Harvard Medical School, Boston, MA
Lexical Technology, Inc., Alameda, CA

Problem lists assist in organizing patient information in computer based medical records. However, in order to use problem lists for billing, research, decision support and standardization, a categorization of the problems entered is required. We describe the problem list component of our computerized patient record, the On-line Medical Record (OMR), which combines a free-text entry mechanism with a categorization scheme, using a dictionary containing 846 terms. All 118,040 problems entered during the system's six years of use have been analyzed, 477 clinicians have entered a mean±S.D. of 238±604 problems into 22,311 patient records. The average number of problems in each patient's file was 5.1±3.9. Comments were typed for 80,281 (68%) of the problems, ranging in length from 1 to 2456 characters, with a mean length of 98±110 characters. Half the problems were entered on the day of the encounter with the patient. Overall, 66% of all problems were categorized in relation to terms from the problem dictionary. Lexical analysis of all problem names showed that 80% could be mapped to Meta 1.4, Snomed 3.0 or a pre-release version of Read 3.0.

We conclude that a problem list entry scheme combining free-text entry and optional categorization using a dictionary can result in a high proportion of problems being categorized as desired. Improvement of the system by elimination of unused dictionary terms and addition of 1000 terms identified by the lexical analysis is likely to result in even higher categorization rates.

The problem list has been identified as one of the key elements of the electronic medical record. Problem lists are short descriptions of past and present medical problems, usually written by clinicians themselves, which are thought to provide the best available summary of a patient's medical condition. The information collected in problem lists can be used for decision support, creation of medical documentation such as discharge summaries, reimbursement and billing information, and research. Most importantly, when uniform terminology is used to categorize each entry, problem lists offer a method for standardization of the medical record.

Vocabularies and coding schemes such as UMLS, ICD-9, Snomed and Read have been suggested to categorize entries to problem lists. However, these standardized lists have been found not to be comprehensive enough to permit a clinically accurate description of patients problems. A poor representation of problem lists phrases by ICD-9 has been demonstrated (1). Evaluation of all four systems for completeness in representation of patients problems indicated that no scheme could be considered comprehensive (2). Lists such as UMLS were found to have incomplete coverage of terms in the fields of hypertension, radiology and ambulatory medicine (3,4,5).

On the other hand, use of free text prevents categorization and most of the perceived benefits listed above. It also enables clinicians to record patient-specific information that needs to be shared among care providers but might never be anticipated in a formal vocabulary, such as depression due to a loss of a pet, for example. In our experience, clinicians' control of the terms used to describe their patients' problems is essential for the widespread acceptance and use of the OMR (6). For example, many clinicians reject the label "hypertension" and prefer to use the expression "elevated blood pressure".

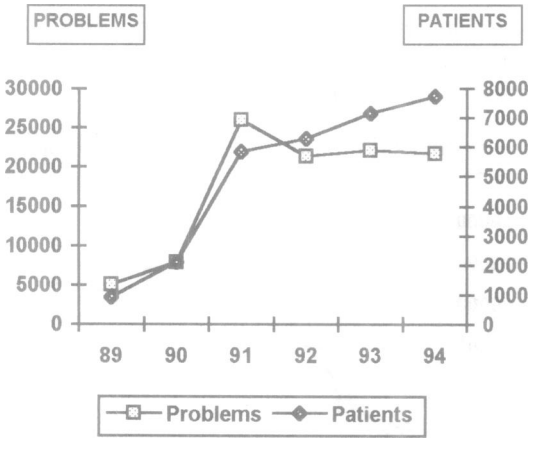
Solutions such as short picklists of frequently used diagnoses (7) and systems permitting un-categorical entries (8) along with categorized entries have been suggested as means to enhance problem capture. However,

0195-4210/95/$5.00

little information exists about acceptance of such systems and their success in categorizing the problems entered.

In this article, we describe the problem list component of the OMR, used at Healthcare Associates (HCA), the academic general internal medicine outpatient clinic of Boston's Beth Israel Hospital (6). This system captures free-text entries as problem names and provides clinicians with suggested categorization for those names, without a change in content. By doing so, this system combines a free text data entry mechanism with an effective categorization scheme, resulting in a high proportion of problems being categorized.

Figure 1. *Annual number of patients and problems in the OMR at HCA clinics, 1989-94*



METHODS

Setting

Problem lists play a central role in the structure and functionality of the OMR, an electronic medical record primarily used in our outpatient clinics.

The OMR was developed at the Beth Israel Hospital by the Center for Clinical Computing, as an extension of the hospital's clinical information system. This system has evolved since the late 1970's to become a comprehensive hospital wide computing environment, providing the hospital's personnel with advanced administrative, communications, decision support and clinical data repository services (9). The system is based on a minicomputer network with more than 2000 terminals throughout the hospital, and uses a combination of hierarchical and relational databases as a programming environment. The OMR was introduced to the General Internal Medicine and Primary Care clinics, Healthcare Associates, in 1989 (6). Since 1993, the use of OMR has been extended into other clinics around the hospital. Nowadays, The OMR supports providers treating patients at 14 of Beth Israel Hospital's outpatient clinics. The OMR includes medication sheets, progress notes, letters and telephone contacts and flow sheets, as well as administrative and laboratory data.

Design

The problem list is used as the navigational tool in the OMR system. Problems can be entered into the OMR as standalone entities, but preferably they are linked to some context. New problems can be added to the problem list at any time. The user is prompted to enter the problem name as free text. A dictionary containing 846 categories is then used to classify the entries. Each category has one preferred name, but synonyms, abbreviations and names with similar meanings can also be used to identify and classify entries. Altogether, there are 1271 entries in the dictionary. After the clinician types a problem name, the computer searches for it in the dictionary, and displays all matches. The user is prompted to choose a match. If no matches are found in the dictionary, or if the user does not choose a match from the list presented, then she or he is given the choice of entering the text as a non-dictionary entity or using another name. The starting date of the problem and its current status (active versus inactive) can be entered. A comment, composed of unstructured free text of unlimited length, can be added to each problem.

Once a problem has been entered into the record, it can be viewed either as part of a standalone problem list screen or as part of a summary screen, together with the patient's currently prescribed medications and recent appointments at the hospital clinics. The provider can edit any element of the problem or delete the problem from the list. Therefore, there are two ways to inactivate a problem : either by choosing the "Inactive" attribute in the problem status field or by simply deleting the problem.

Problems can be linked to content each time a progress note is entered in the OMR. At the end of the note editing process, the note writer is

offered a chance to link the note, by order of relevance, to the problems already in the patient's problems list. New problems that need to be linked to the new note can be added as well. Afterwards, the note is presented in the note list with the first problem it is linked to. The note reader can choose to view only notes that have been linked to a specific problem, which is useful as a filtering mechanism when one is reading notes in a record with multiple notes.

Another method of linking problem names to content is through an expert system. The problem list is used to identify new HIV-positive patients. Once such a patient has been identified, his or her providers can use a set of decision support sources, data entry mechanisms, and reminders and alerts based on the current guidelines for treatment of ambulatory HIV-positive patients.

Data Collection and Analysis

All the problems stored in the OMR database since it was introduced in 1989 have been searched, and the data were analyzed with the SAS package Release 6.09 for statistical analysis (SAS Institute, Cary, NC).

The textual content of all problem names and categories entered during 1994 were analyzed with a lexical parsing program (10). The text entered was compared for matches in three coding schemes: Meta 1.4 - the current version of the UMLS Metathesaurus; Snomed International Version 3.0; and a pre-release version of Read (Version 3.0).

RESULTS

Problems

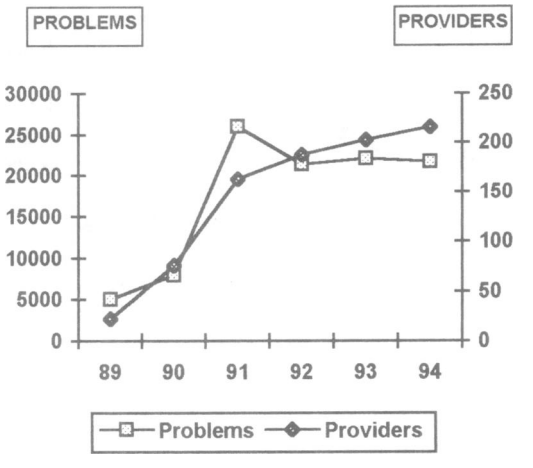
In the period from the introduction of OMR in 1989 until March 1995, 118,040 problems were entered for 22,311 patients. Figure 1 shows the annual number of problems and the number of patients for whom new problems have been entered since the system was introduced. The mean number of problems per patient was 5.1±3.9 with a range of 1-34 and a median of 5.

Comments were added to 80,281 (68%) of the problems, ranging in length from 1 to 2456 characters. The mean length of the comments (±SD) was 98±110 characters, with a median of 63 characters. Most of the longer comments dated to the early development stages of the OMR, when notes were not available and some providers used the comments option as a method of capturing data about encounters with patients.

Providers

Four hundred seventy-seven health care providers including physicians with different levels of training, nurses, resource specialists and social workers have entered problems. Of these, 388 providers, each entering more than five problems, have entered 98.9% of all existing problems. The annual number of providers and problems entered is shown in Figure 2. The mean number of problems entered by each provider was 238±604, with a range of 1 to 5953. At the General Internal Medicine clinic, 24 providers have entered 50% of all the problems, and 116 providers have entered 80%.

Figure 2. *Annual number of providers entering problems at HCA clinics, 1989-94 (with annual problems curve from figure 1 repeated for comparison).*



Of the 118,040 problems, 10,351 (8.8%) were inactive and 4480 (3.8%) were deleted. The remaining problems were classified as active.

Problem Categories

Overall, 66% of all the problems entered have been categorized as relating to one of the 846 dictionary terms. Analysis of the types of categorized problems shows that 40 dictionary terms constitute 50% of all the user-categorized problems occurring in our data set. A list of the

first 15 most frequent problems is presented in Table 1. In 1994, 240 out of the 846 categories in the dictionary were not used even once for categorization of entered problem names.

Table 1. *List of the 15 most frequently used diagnosis categories in OMR.*

Category Name	Occurrence
Health Maintenance	4863
Hypertension	4698
Psychosocial	2524
Tobacco abuse	1868
Hypercholesterolemia	1827
Obesity	1655
Depression	1553
Asthma	1234
Headache	1147
CAD	974
Anemia	959
Allergy	827
Low Back Pain	792
PPD Positive	721
Back pain	717

Timing of Problem Entry

In 1994, there were 39,144 visits of 12,499 patients to the General Internal Medicine clinic. During that period, 22,240 problems were entered for these patients. Of these problems, 11,196 (50.3%) were entered on the day of visit. We presume that a large fraction of the remaining problems were entered at the time transcribed notes were electronically signed, within a few days of a visit.

Lexical Analysis

Problem names for 1994 in the General Internal Medicine clinic included 15,171 (68.2%) categorized problem names and 7079 (31.8%) non-categorized names. All the categorized names have matches to ICD-9 codes.

From the categorized names, 13,433 (88.5%) were mapped to terms in Meta 1.4, Snomed International 3.0 or a pre-release version of Read 3.0. Of the 7079 non-categorized terms, 2650 (37.4%) could be mapped to one of these dictionaries. Overall, 80% of the problems entered in 1994 could be mapped to existing dictionaries.

DISCUSSION

Limited data are available about the acceptance, usability and evaluation of different methods for capturing problem lists into computerized medical record systems. In this study, we have shown that our system has succeeded, after a two-year adaptation period, in achieving a high and stable level of usability. This is shown by the large numbers of clinicians using the system for the daily care of their patients, and by the large numbers of problems entered.

The early and close timing of problem entry near the patient's visit and the relatively long comments, entered by the clinicians themselves, suggest that the clinicians value the information stored in the computerized record and are willing to invest the time required for data entry to obtain the future benefits offered by this system. Such extensive use is evidence for wide acceptance of the clinical computing system and also for the helpfulness of the problem lists in the process of patient care. Most clinicians use the problem lists during patients appointments for structuring the encounter and to create structured progress notes.

Wilton (8) has studied 2903 problems entered for 3385 pediatric patients. The problems were entered with a system that allows free-text entry but prompts the user with problem names from a predefined list of 328 coded problems names and abbreviations. He found that 82% of the problems in his database were selected from an on-screen list of 328 common problems, and another 15.4% were found later in a more comprehensive database of ICD-9 codes, bringing the number of categorized problems to 97.4%. However, in his system, data were not entered directly by clinicians but were transcribed from encounter forms, and the exact vocabulary used by the clinician was not preserved in 82% of cases, but was selected from a predefined list. With our system 80% of all problems were categorized for a much larger sample of patients and problems and over a longer period of time, while maintaining the exact terms used by the clinician when the problems were entered.

Our study has shown that entry of problem names as free text is widely accepted by a large number of clinicians. Clinicians have been shown not to be satisfied with the currently existing coding schemes, such as ICD-9 (1), UMLS, Snomed and Read (2). That made it

necessary to add to lists of names prepared for use as standard vocabularies for problem lists some "local" phrases, which do not exist even in extensive lists such as the Metathesaurus, but were demanded by clinicians (5). The large dictionaries that are in use, such as the UMLS Metathesaurus, are not permanent lists of terms. Because of a steady evolution in the medical vocabulary, an estimated 10% of all terms undergo change each year (11). For all these reasons, a solution that is not dependent on a rigid dictionary of terms but still allows linking to a predefined list of definitions is preferable.

Creating a computerized problem list that simultaneously describes the patient's problems using the clinician's vocabulary, and allows a full categorization of the problems still remains a challenge. However, we feel that our system is getting close to achieving that goal. Improvements that can be introduced into the system as a result of this study are enhancements to the dictionary. The addition of more than 1000 terms that were suggested by the lexical analysis of our non-matchables could increase matching with Meta 1.4 to well over 90%. However, we speculate that an 80% categorization level is sufficient for decision support and outcomes research. We still need to find ways of increasing the rate of editing and updating of the problem lists. Further analysis of nonspecific problem names such as "health maintenance" and "psychosocial" is required. Are these nonspecific terms used as measures for confidentiality, or are they used because the clinician wants to communicate problem list entries with the patient, and therefore prefers not to use more explicit terms?

The free-text categorized problem list has enabled providers at our clinics to capture and categorize large numbers of problem names without compromising the accuracy of the terms used or limiting the vocabulary. However, such a system has little use without the additional parts of a comprehensive computerized records: progress notes, medications, flow sheets and decision support. By integrating the problem list with the rest of the medical record, a high level of functionality, and a high level of user satisfaction, can be maintained. We hope that further improvement of our system will permit further integration and promote usability in other outpatient clinics in the hospital as well as in community primary care centers affiliated with the hospital.

Supported, in part, by cooperative agreement number HS-08749-01 from the National Library of Medicine and the Agency for Health Care Policy and Research, and by research funds from the Center for Clinical Computing

---

1. Payne TH, Murphy GR, Salazar AA. How well does ICD-9 represent phrases used in the medical problem list. Proc 17th Ann Symp Comput Applic Med Care 1993; 654-657.
2. Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. Proc 18th Ann Symp Comput Applic Med Care 1994; 201-205.
3. Campbell JR, Kallenberg GA, Sherrick RC. The clinical utility of Meta: an analysis for hypertension. Proc 16th Ann Symp Comput Applic Med Care 1992; 397-401.
4. Friedman C. The UMLS coverage of clinical radiology. Proc 16th Ann Symp Comput Applic Med Care 1992; 309-313.
5. Payne TH, Martin DR. How useful is the UMLS Metathesaurus in developing a controlled vocabulary for an automated problem list. Proc 18th Ann Symp Comput Applic Med Care 1994. 705-709.
6. Safran C, Rury C, Rind DM, Taylor WC. A computer based outpatient medical record for a teaching hospital. MD Computing 1991; 8(5):291-9.
7. Scherpbier HJ, Abrams RS, Roth HR, Hail JJB. A simple approach to physician entry of patient problem lists. Proc 18th Ann Symp Comput Applic Med Care 1994; 206-210.
8. Wilton R. Non categorical problem lists in a primary care information system. Proc 15th Ann Symp Comput Applic Med Care 1991; 823-27.
9. Bleich HL, Beckley RF, Horowitz GL, Jackson JD, Moody ES, Franklin C, Goodman SR, McKay MW, Pope RA, Walden T, Bloom SM, Slack WV. Clinical Computing in a teaching hospital. NEJM 1985; 312:756-764.
10. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc 18th Ann Symp Comput Applic Med Care 1994; 235-239.
11. Tuttle MS, Shererez DD, Erlbaum MS, Sperzel WD, Fuller LF, Olson NE. Adding your terms and relationships to the UMLS Metathesaurus. Proc 16th Ann Symp Comput Applic Med Care 1992. 219-223.